

# DØ Computing Experience and Plans for SAM-Grid

EU DataGrid

Internal Project Conference

May 12-15, 2003

Barcelona

Lee Lueking

Fermilab

Computing Division

## Roadmap of Talk

- DØ overview
- Computing Architecture
- SAM at DØ
- SAM-Grid
- Regional Computing Strategy
- Summary

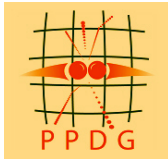


May 12-15, 2003

Lee Lueking, EDG Int. Proj. Conf.



# The DØ Experiment

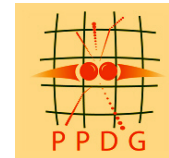


- DØ Collaboration
  - 18 Countries; 80 institutions
  - >600 Physicists
- Detector Data (Run 2a end mid '04)
  - 1,000,000 Channels
  - Event size 250KB
  - Event rate 25 Hz avg.
  - Est. 2 year data totals (incl. Processing and analysis):  $1 \times 10^9$  events,  $\sim 1.2$  PB
- Monte Carlo Data (Run 2a)
  - 6 remote processing centers
  - Estimate  $\sim 0.3$  PB.
- Run 2b, starting 2005:  $>1$  PB/year



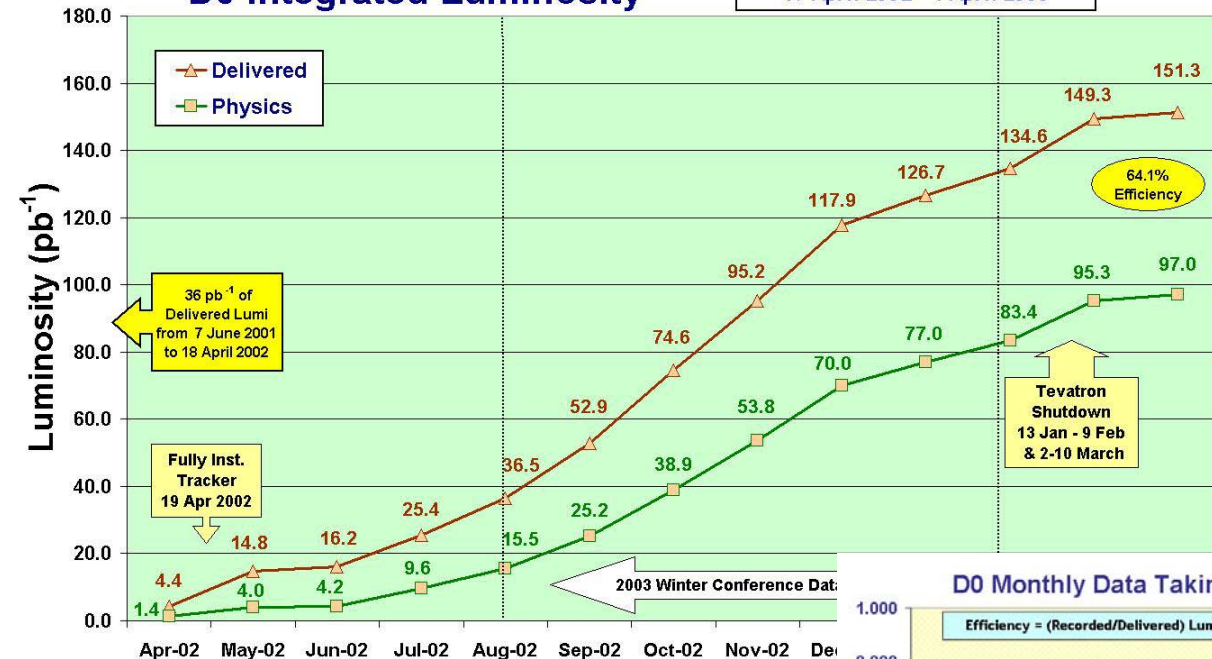


# DØ Experiment Progress



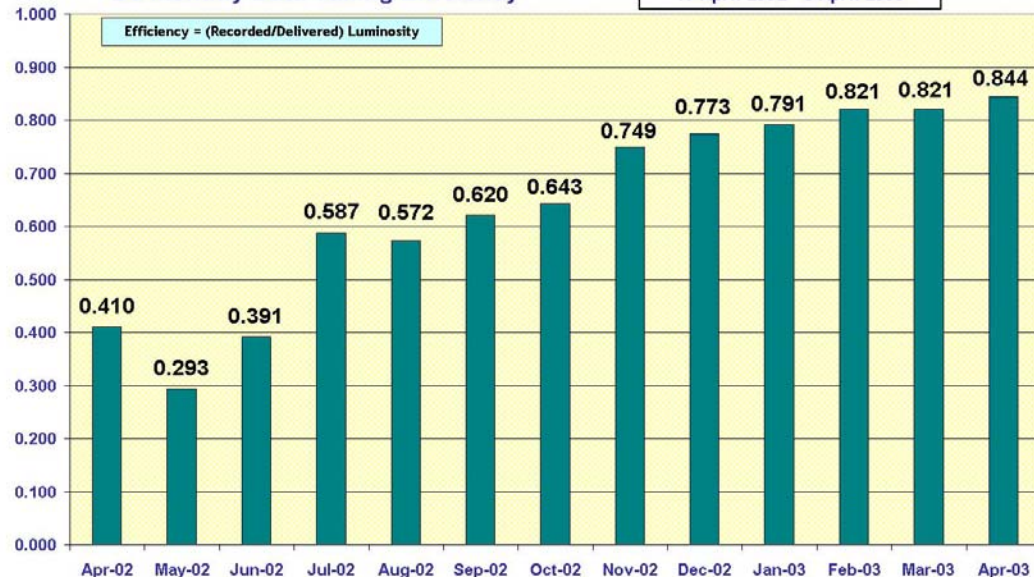
## DØ Integrated Luminosity

19 April 2002 - 4 April 2003



## DØ Monthly Data Taking Efficiency

19 April 2002 - 8 April 2003



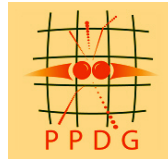
May 12-15, 2003

Lee Luek

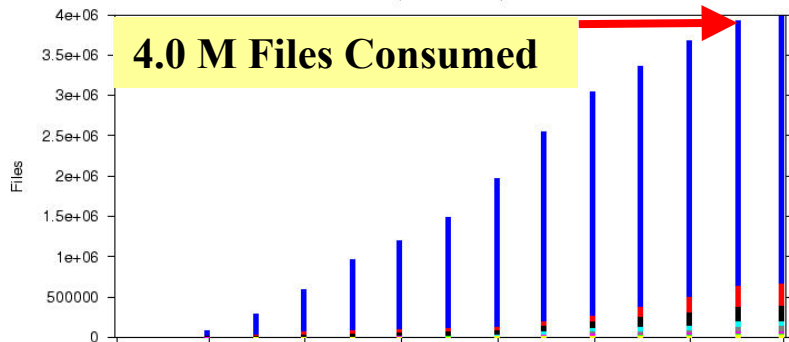




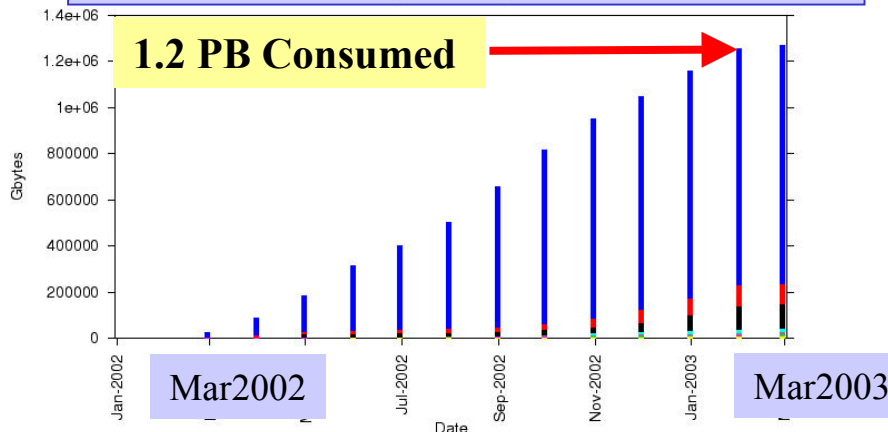
# Overview of DØ Data Handling



## Integrated Files Consumed vs Month (DØ)



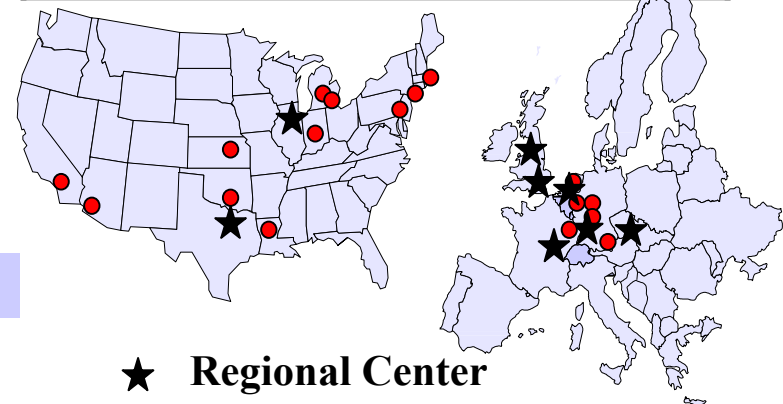
## Integrated GB Consumed vs Month (DØ)



Station  
 central-analysis  
 final-farm  
 cab  
 clued0  
 d0karlsruhe  
 imperial-test  
 triviala  
 other

## Summary of DØ Data Handling

Registered Users	600
Number of SAM Stations	56
Registered Nodes	900
Total Disk Cache	40 TB
Number Files - physical	1.2M
Number Files - virtual	0.5M
Robotic Tape Storage	305 TB



★ **Regional Center**  
 ● **Analysis site**

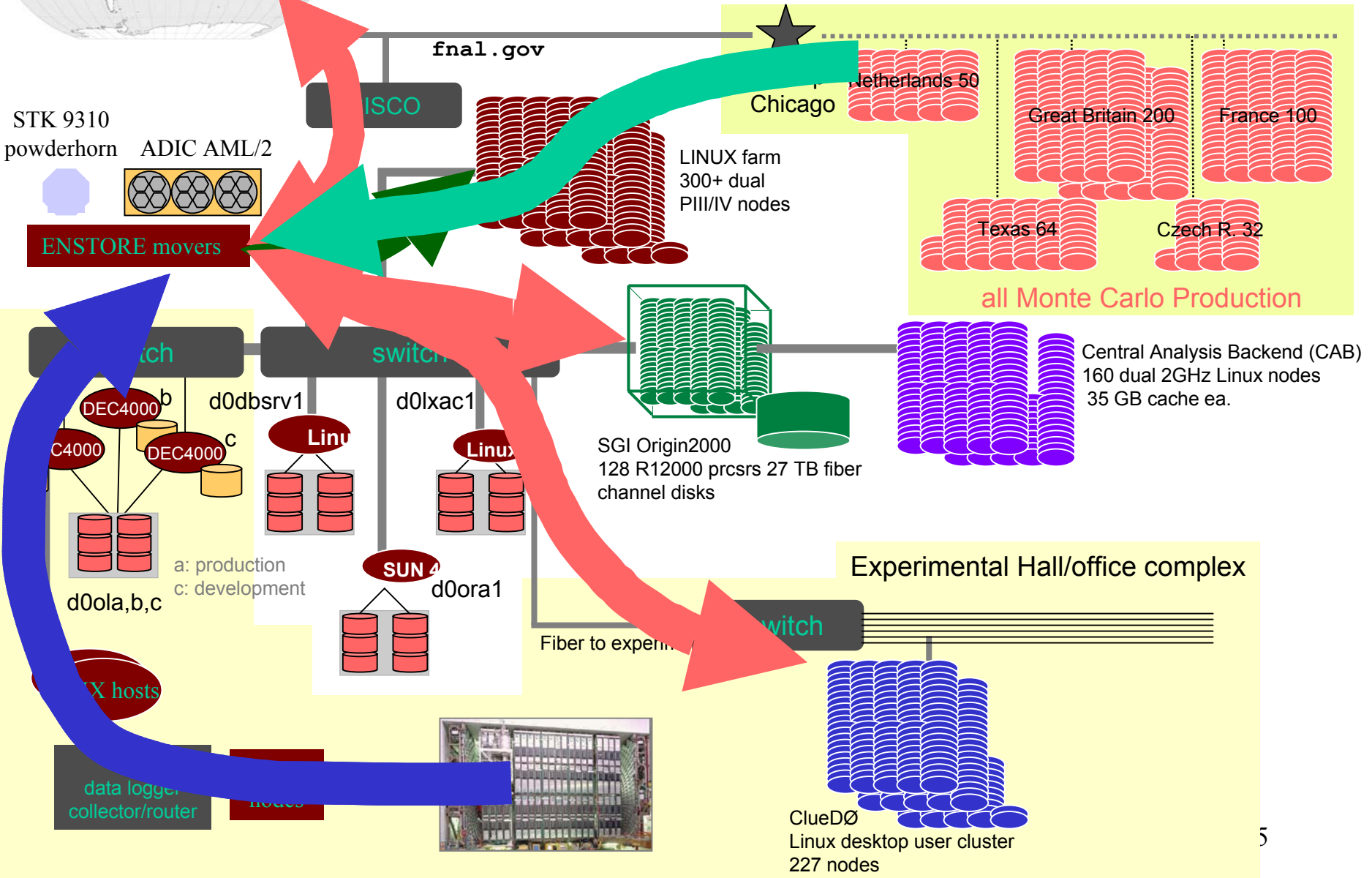
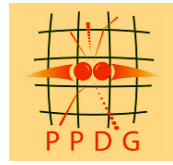
May 12-15, 2003

Lee Lueking, EDG Int. Proj. Conf.





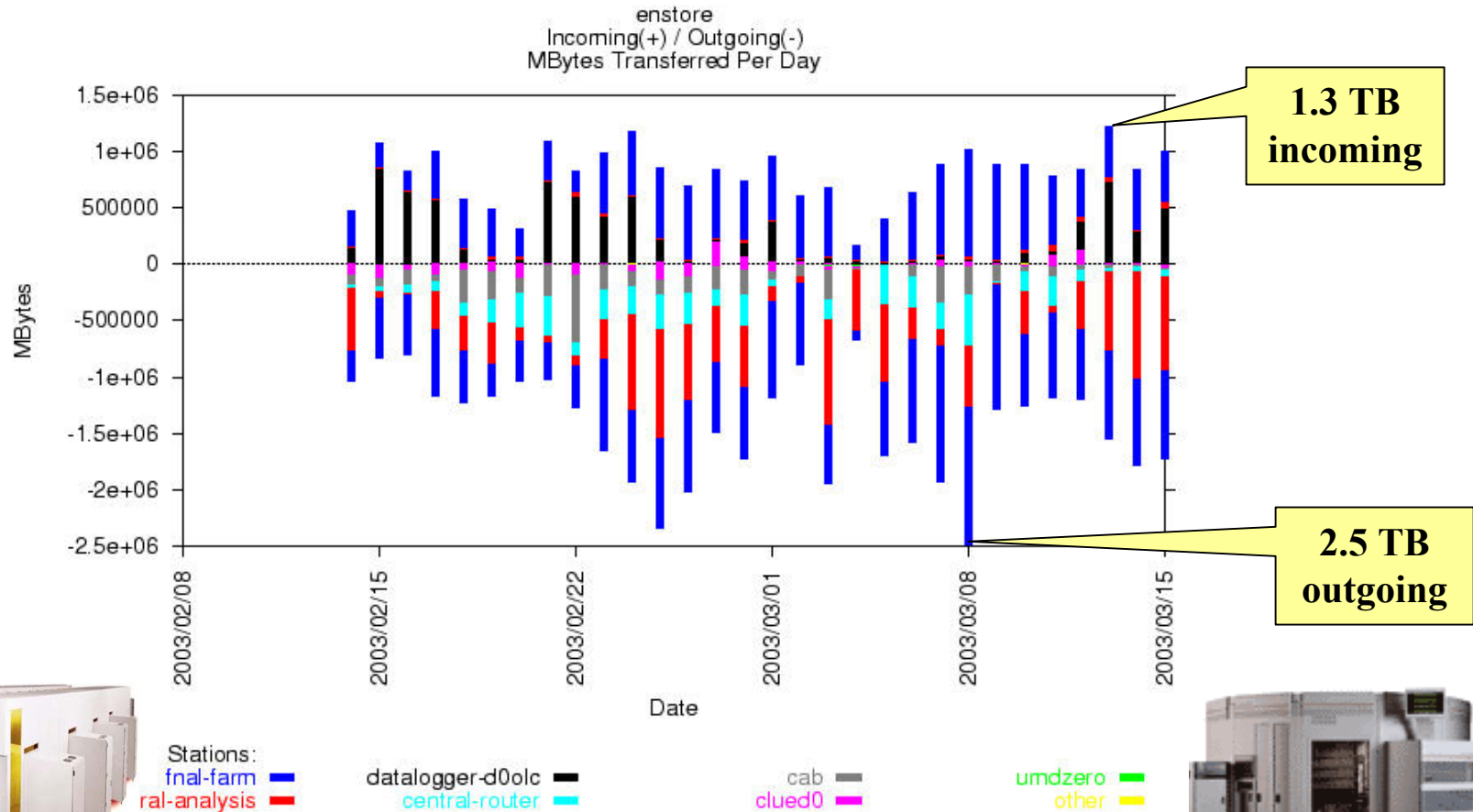
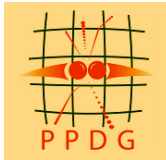
# DØ computing/data handling/database architecture

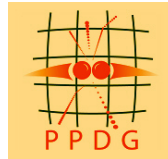




# Data In and out of Enstore

(robotic tape storage) Daily Feb 14 to Mar 15





## SAM at DØ



[d0db.fnal.gov/sam](http://d0db.fnal.gov/sam)

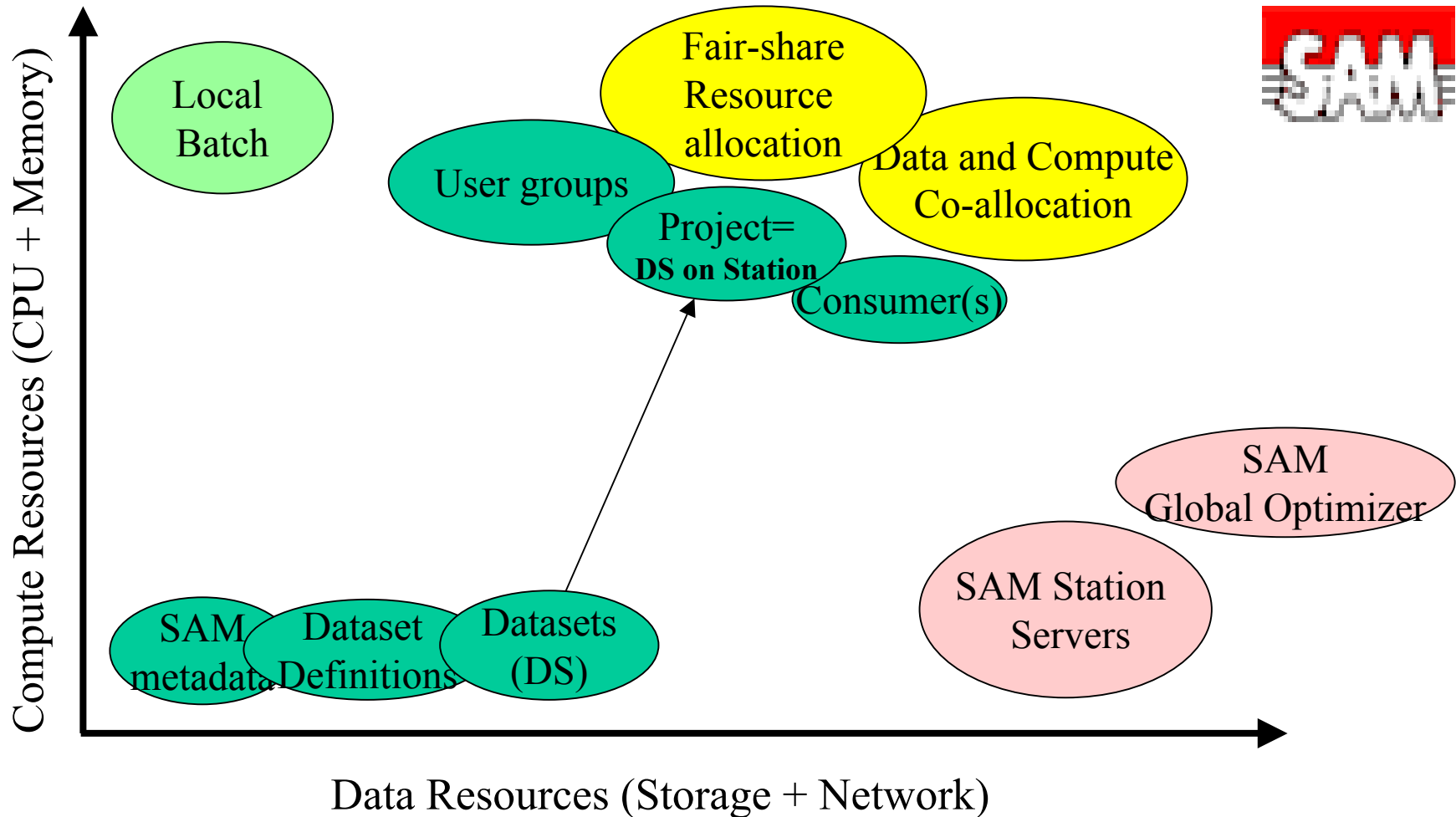
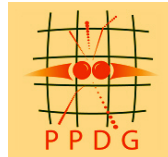


May 12-15, 2003

Lee Lueking, EDG Int. Proj. Conf.

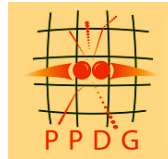


# Managing Resources in SAM





# SAM Features

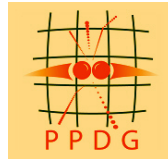


- **Flexible and scalable model**
- **Field hardened code**
- **Reliable and Fault Tolerant**
- **Adapters for many local batch systems: LSF, PBS, Condor, FBS**
- **Adapters for mass storage systems: Enstore (FNAL), HPSS (Lyon), and TSM (GridKa)**
- **Adapters for Transfer Protocols: cp, rcp, scp, encp, bbftp, GridFTP**
- **Useful in many cluster computing environments: SMP w/ compute servers, Desktop, private network (PN), NFS shared disk,...**
- **User interfaces for storing, accessing, and logically organizing data**



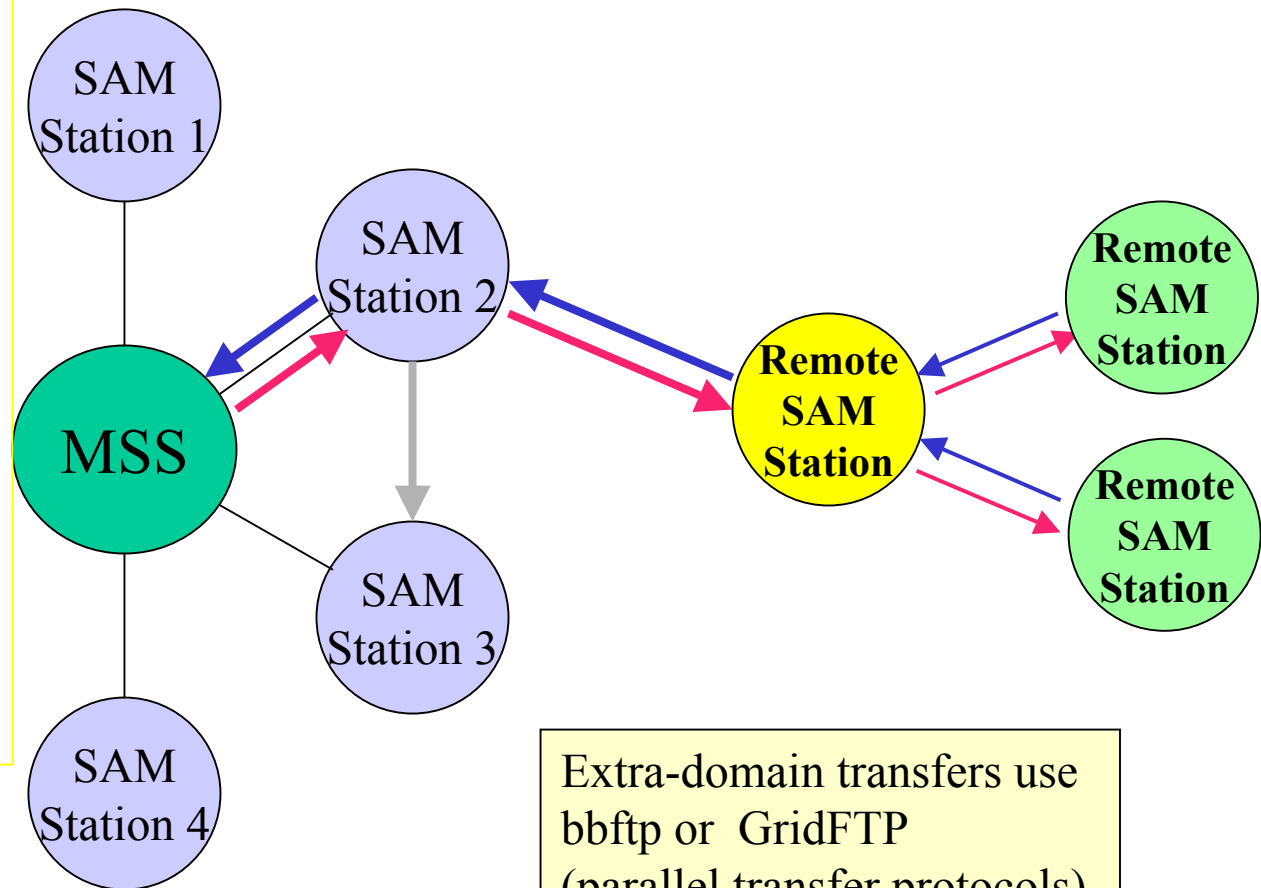


# The SAM *Station* Concept



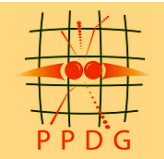
## Station Responsibilities

- Pre-stage files for *consumers*.
- Manage local cache
- Store files for *producers*
- Forwarding
  - File stores can be forwarded through other stations
- Routing
  - Routes for file transfers are configurable





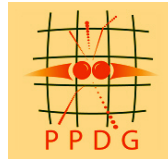
# DØ SAM Station Summary



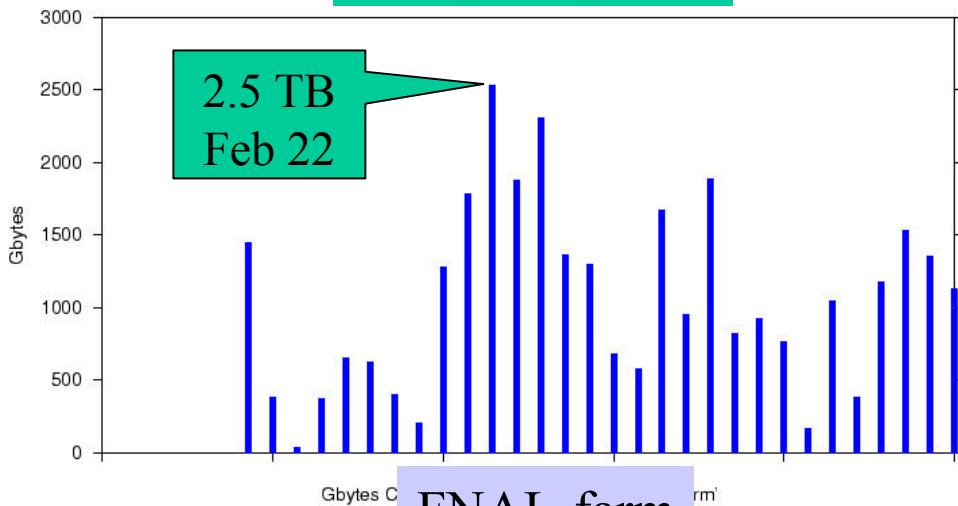
Name	Location	Nodes/cpu	Cache	Use/comments
Central-analysis	FNAL	128 SMP*, SGI Origin 2000	14 TB	Analysis & DØ code development
CAB (CA Backend)	FNAL	16 dual 1 GHz + 160 dual 1.8 GHz	6.2 TB	Analysis and general purpose
FNAL-Farm	FNAL	100 dual 0.5-1.0 GHz +240 dual 1.8 GHz	3.2 TB	Reconstruction
CLueDØ	FNAL	50 mixed PIII, AMD. (may grow >200)	2 TB	User desktop, General analysis
DØkarlsruhe (GridKa)	Karlsruhe, Germany	1 dual 1.3 GHz gateway, >160 dual PIII & Xeon	3 TB NFS shared	General/Workers on PN. Shared facility
DØumich (NPACI)	U Mich. Ann Arbor	1 dual 1.8 GHz gateway, 100 x dual AMD XP 1800	1 TB NFS shared	Re-reconstruction. workers on PN. Shared facility
Many Others > 4 dozen	Worldwide	Mostly dual PIII, Xeon, and AMD XP		MC production, gen. analysis, testing



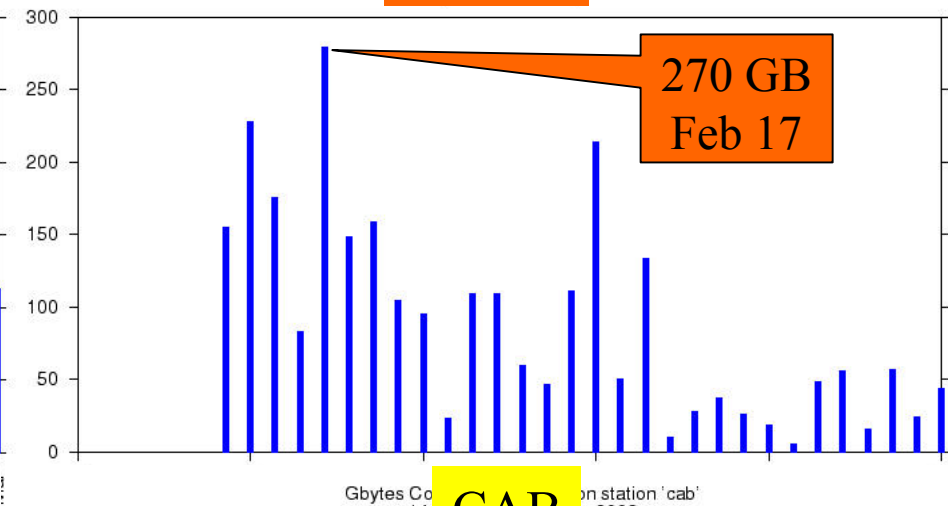
# Station Stats: GB Consumed (by jobs) Daily Feb 14 – Mar 15



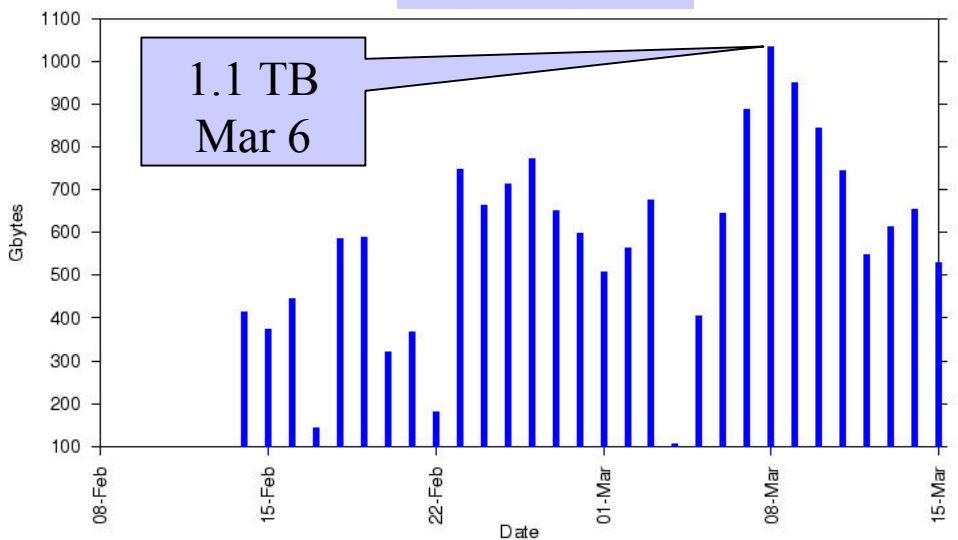
Gbytes Central-Analysis



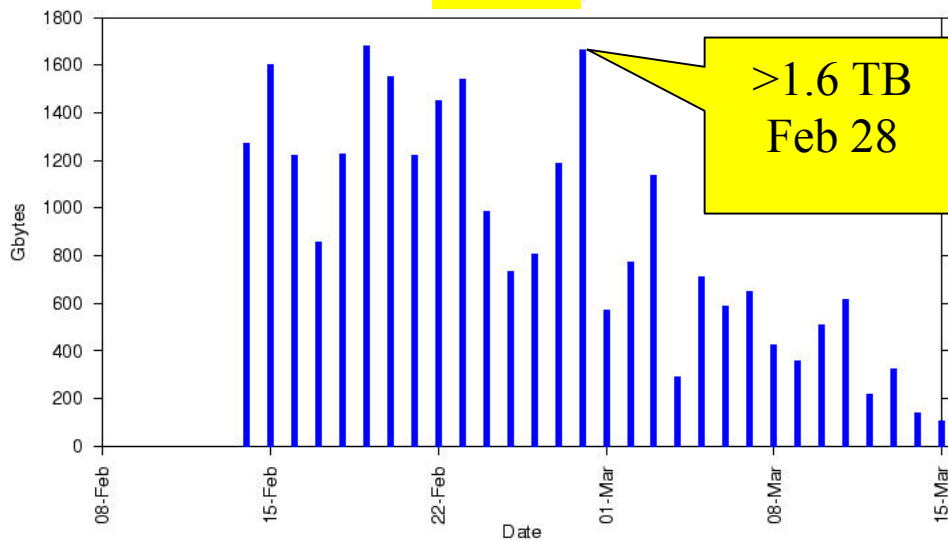
Gbytes ClueD0 station 'clued0' 2003



Gbytes FNAL-farm



Gbytes CAB station 'cab' 2003



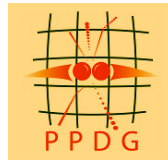
Station  
fnal-farm

Station  
cab

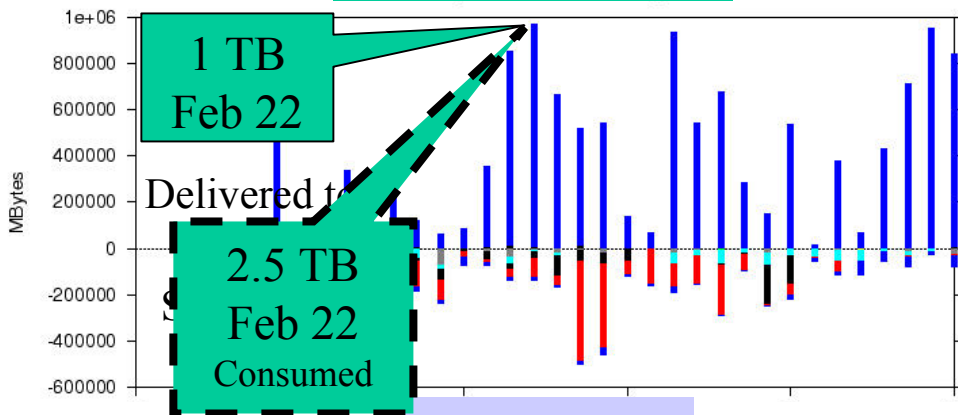


# Station Stats: MB Delivered/Sent

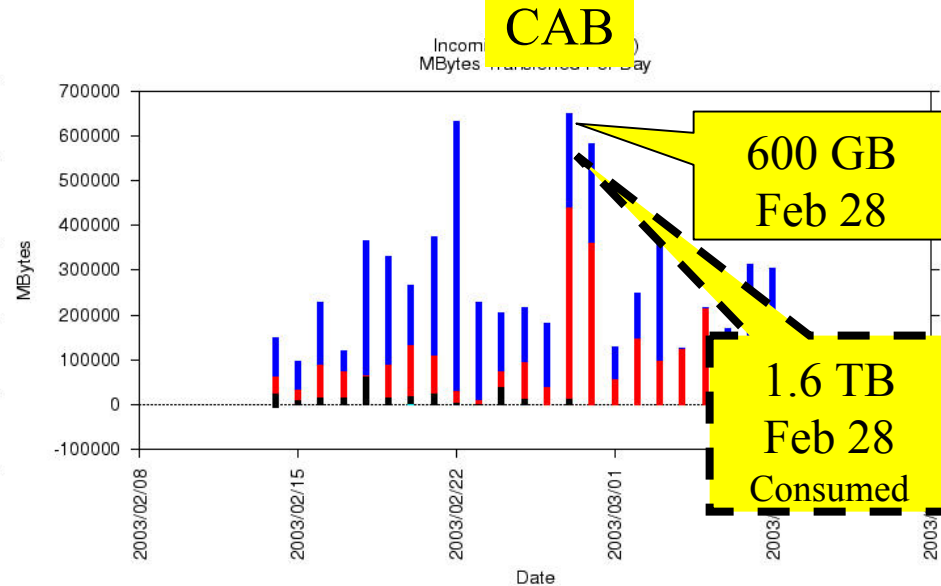
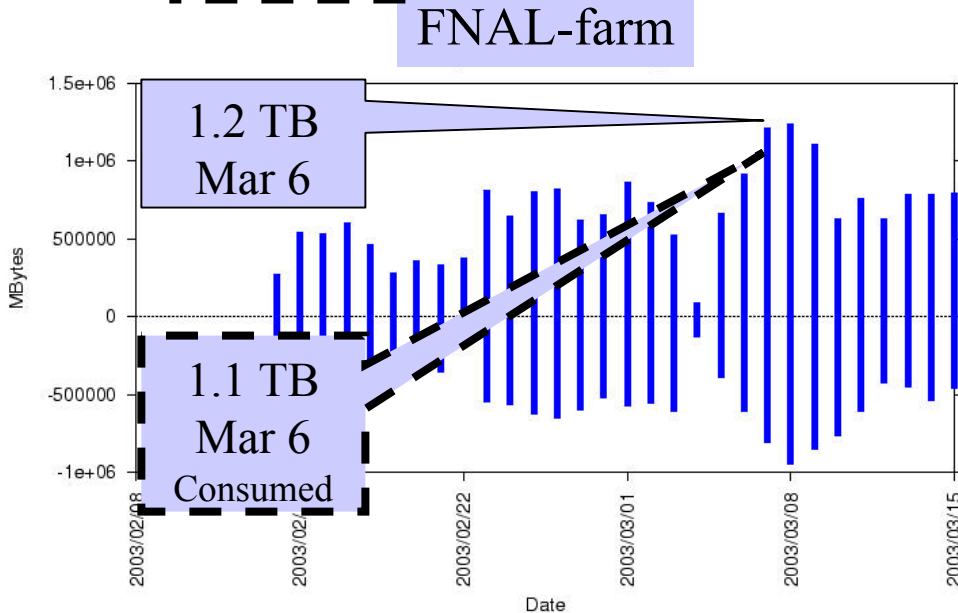
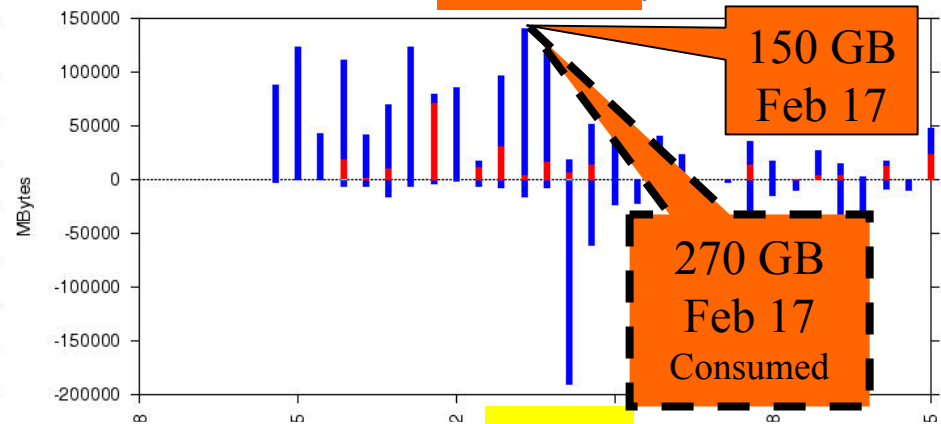
Daily Feb 14 – March 15



## Central-Analysis



## ClueD0



Stations:  
enstore

clued0

Stations:  
central-analysis

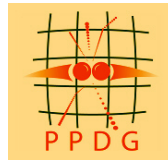
central-router  
princeton-d0

uta-hep  
imperial-test

d0karlsruhe  
umdzero

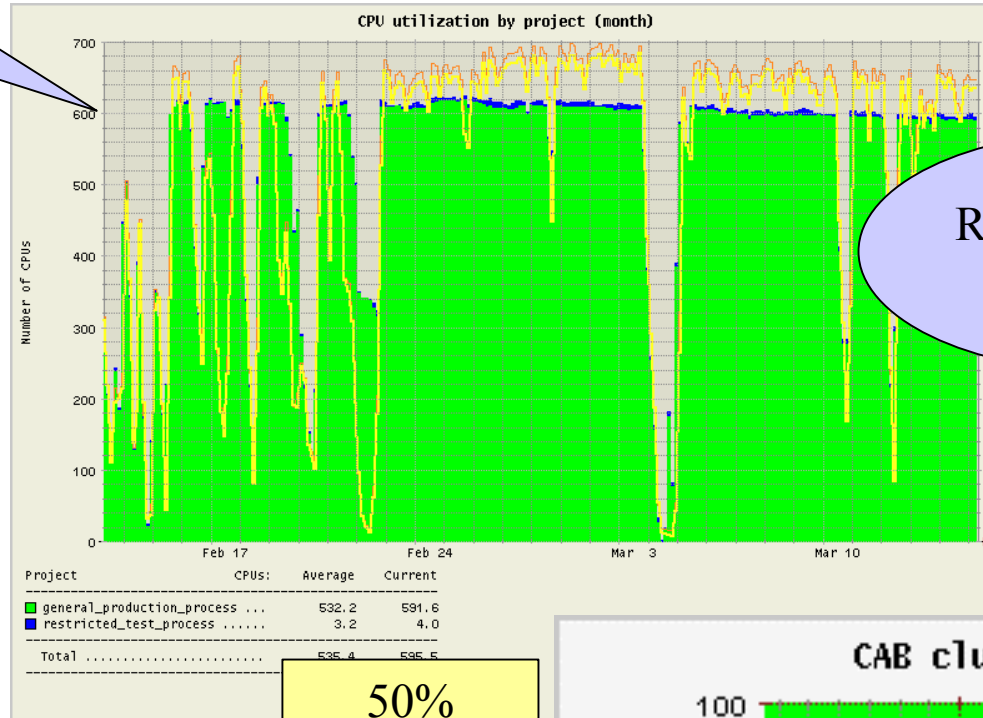


# FNAL-farm Station and CAB CPU Utilization



Feb 14 – March 15

600  
CPUs

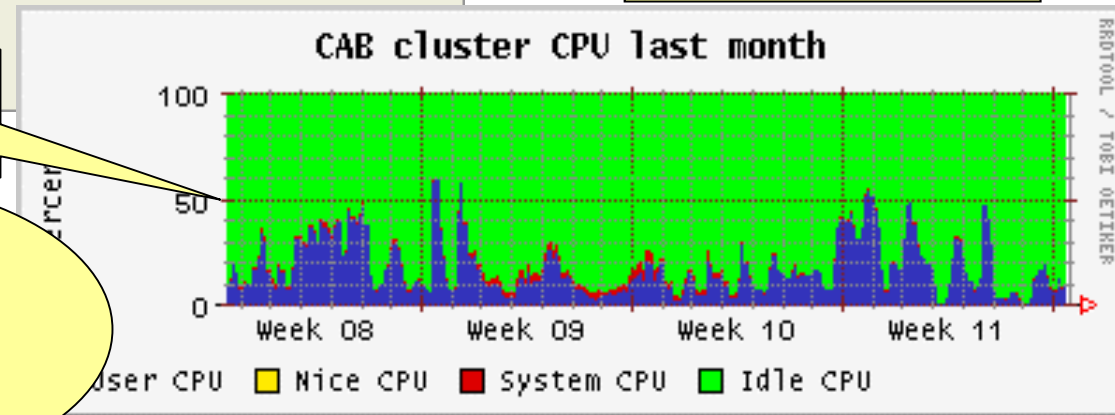


FNAL-farm  
Reconstruction  
Farm  
300 duals

CAB Usage will  
increase dramatically  
in the coming months

50%  
Utilization

Central-Analysis  
Backend  
Compute Servers  
160 duals

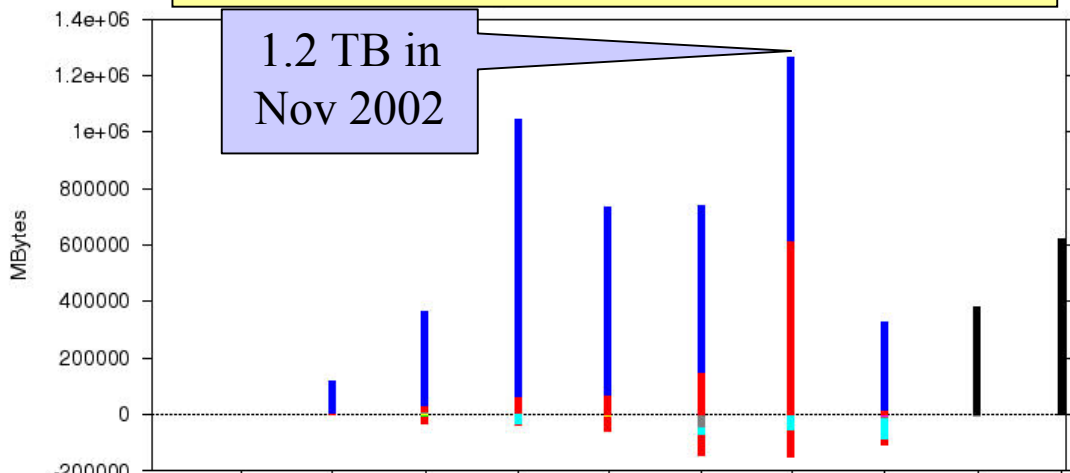




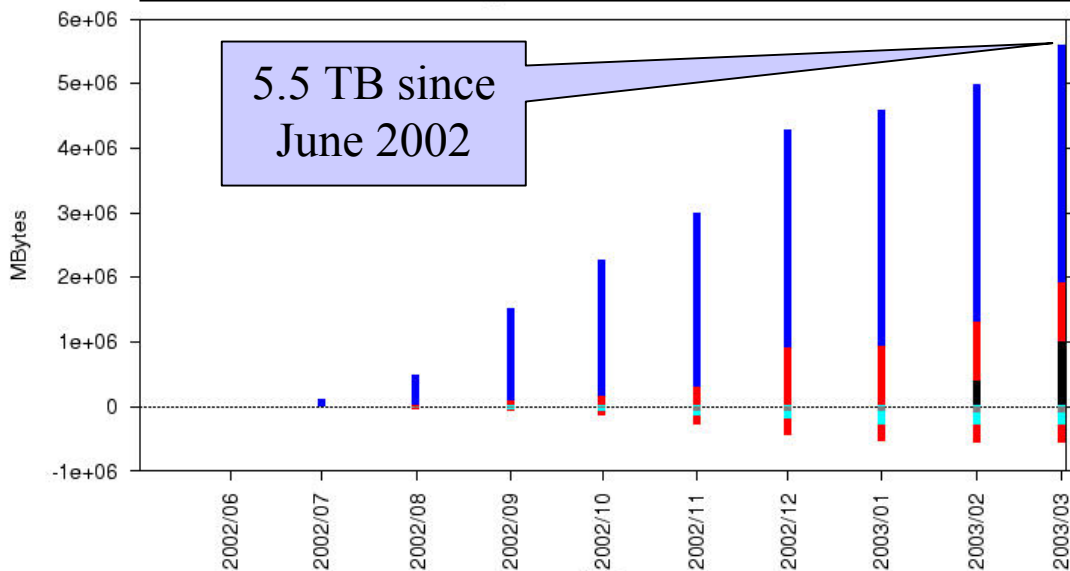
# DØ Karlsruhe Station at GridKa



## Monthly Thumbnail Data Moved to GridKa



## Cumulative Thumbnail Data Moved to GridKa



The GridKa SAM Station uses shared cache config. with workers on a private network

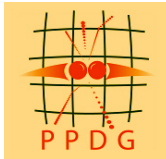


**This is our first Regional Analysis Center (RAC).**

- **Resource Overview:**
  - **Compute:** 95 x dual PIII 1.2GHz, 68 x dual Xeon 2.2 GHz. DØ requested 6%. (updates in April)
  - **Storage:** DØ has 5.2 TB cache. Use of % of ~100TB MSS. (updates in April)
  - **Network:** 100Mb connection available to users.
  - **Configuration:** SAM w/ shared disk cache, private network, firewall restrictions, OpenPBS, Redhat 7.2, k 2.418, DØ software installed.



# Challenges (1)

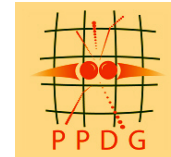


- Getting SAM to meet the needs of DØ in the many configurations is and has been an enormous challenge.
  - **Automating Monte Carlo Production and Cataloging** with MC request system in conjunction with MC RunJob meta system.
  - **File corruption issues**. Solved with CRC.
  - **Preemptive distributed caching** is prone to race conditions and log jams. These have been solved.
  - **Private networks** sometimes require “border” naming services. This is understood.
  - **NFS shared cache configuration** provides additional simplicity and generality, at the price of scalability (star configuration). This works.
  - **Global routing** completed.



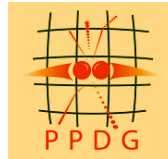


## Challenges (2)



- **Convenient interface** for users to build their own applications. SAM user api is provided for python.
- **Installation procedures** for the station servers have been quite complex. They are improving and we plan to soon have “push button” and even “opportunistic deployment” installs.
- **Lots of details** with opening ports on firewalls, OS configurations, registration of new hardware, and so on.
- **Username clashing issues**. Moving to GSI and Grid Certificates.
- **Interoperability with many MSS**.
- **Network attached files**. Consumer is given file URL and data is delivered to consumer over the network via RFIO, dCap, etc.





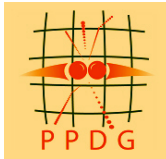
## SAM Grid

<http://www-d0.fnal.gov/computing/grid/>





# DØ Objectives of SAM-Grid

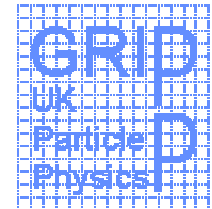


- JIM (Job and Information Management) complements SAM by adding job management and monitoring to data handling.
- Together, JIM + SAM = SAM-Grid
- Bring standard grid technologies (including Globus and Condor) to the Run II experiments.
- Enable globally distributed computing for DØ and CDF.

## •People involved:

–Igor Terekhov (FNAL; JIM Team Lead),  
Gabriele Garzoglio (FNAL), Andrew  
Baranovski (FNAL), Rod Walker (Imperial  
College), Parag Mhashilkar & Vijay Murthi  
(via Contr. w/ UTA CSE), Lee Lueking  
(FNAL; Team rep. For DØ to PPDG)

–Many others at many DØ and CDF sites



**Condor**  
High Throughput Computing



May 12-15, 2003

Lee Lueking, EDG Int. Proj. Conf.

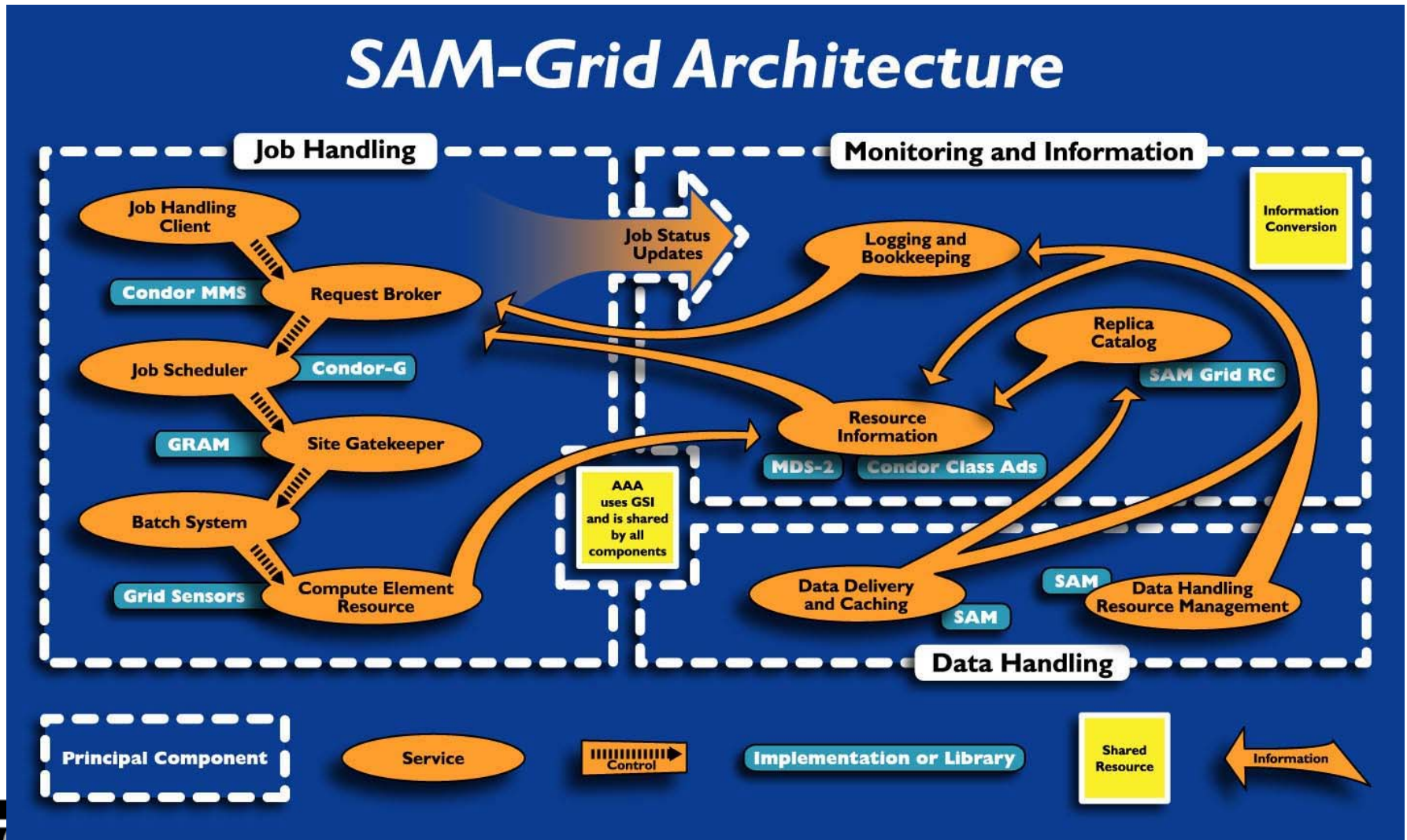
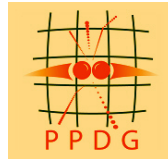


the globus project™

[www.globus.org](http://www.globus.org)

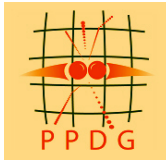


# The SAM-Grid Architecture





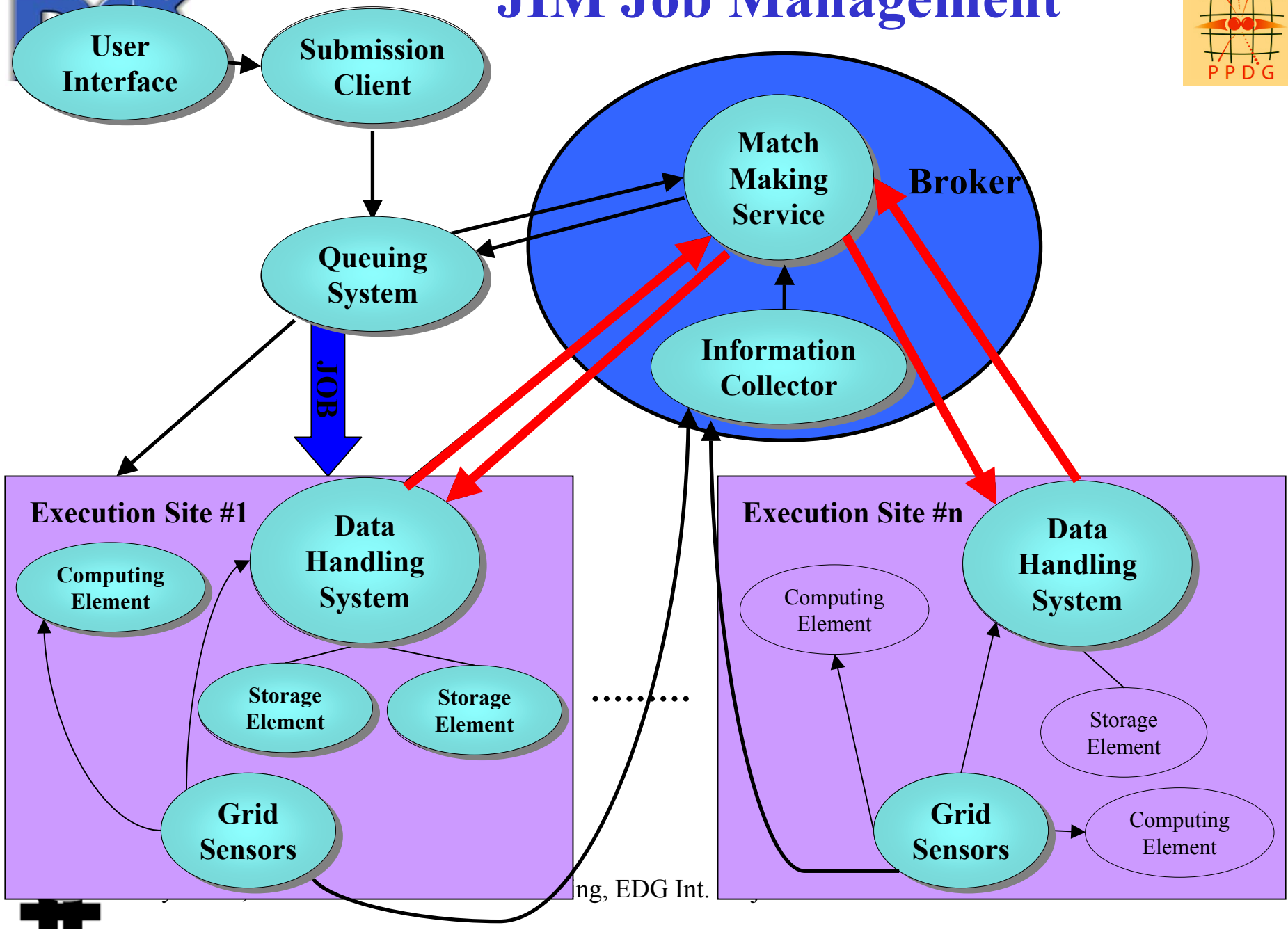
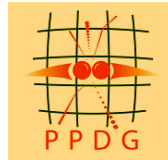
# Condor-G Extensions Driven by JIM



- **The JIM Project team has inspired many Extensions to the Condor software**
  - **Added Match Making to the Condor-G for grid use.**
  - **Extended class adds to have the ability to call external functions from the match making service.**
  - **Introduced a three tier architecture which separates the user submission, job management service, and submission sites completely.**
- **Decision making on the grid is very difficult. The new technology allows:**
  - **Including logic not expressible in class ads**
  - **implementing very complex algorithms to establish ranks for the jobs in the scheduler**
- **Also, many robustness and security issues have been addressed**
  - **TCP replaces UDP for communication among Condor services**
  - **GSI now permeates the Condor-G services, driven by the requirements of the three-tier architecture**
  - **Re-matching a grid job that failed during submission**

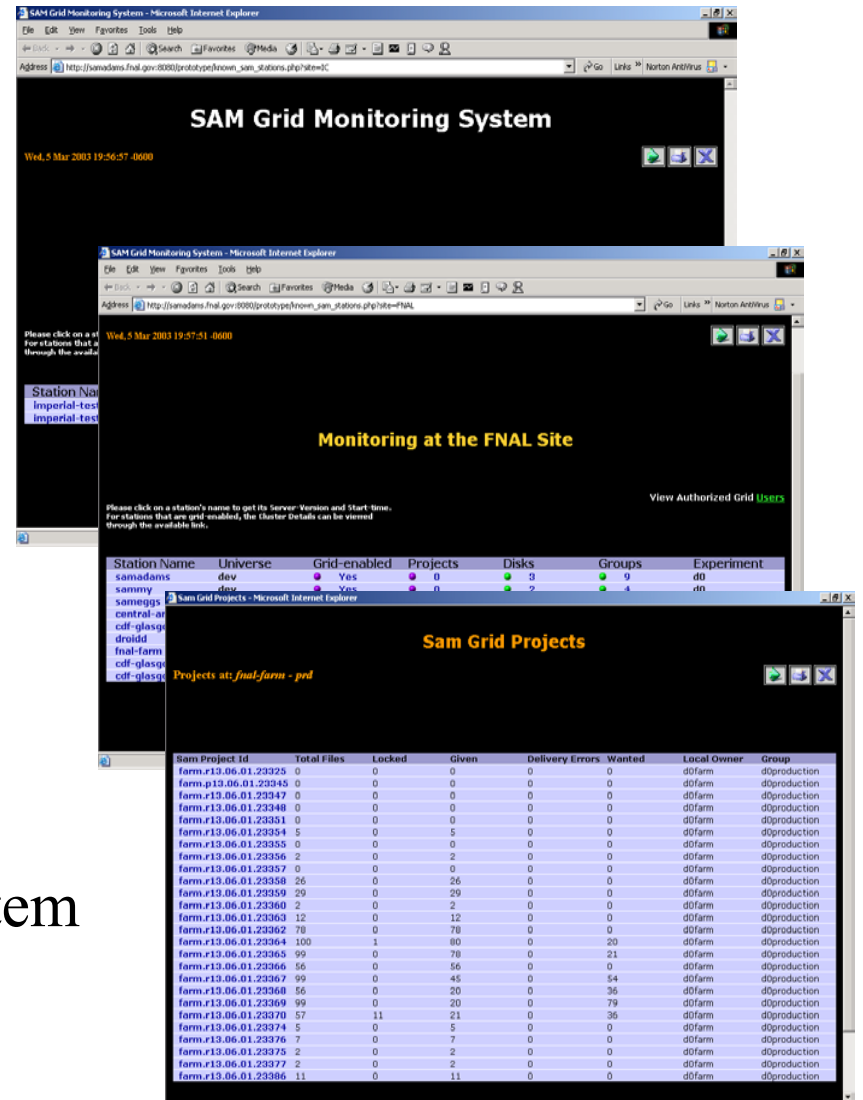
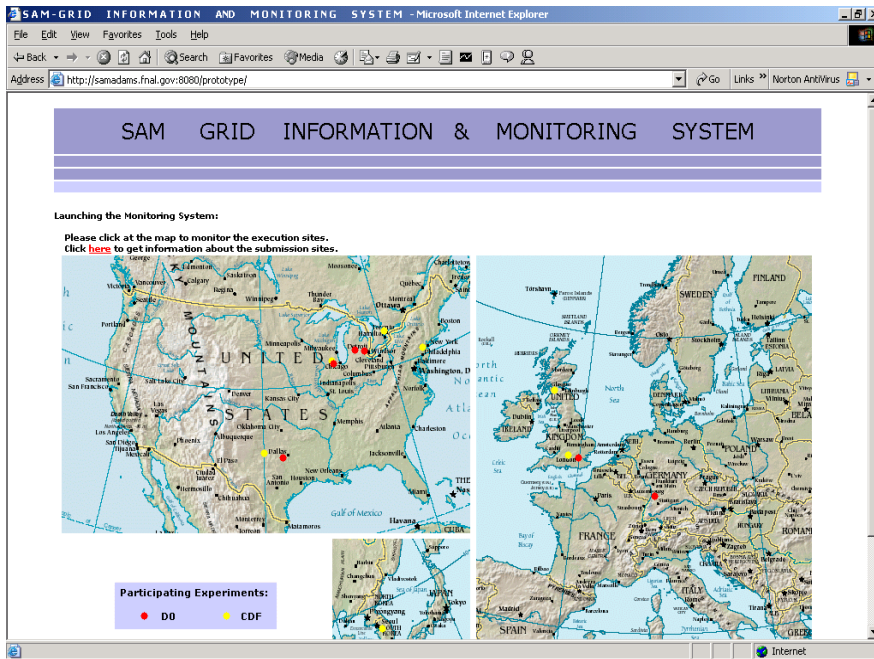
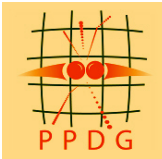


# JIM Job Management





# SAM-Grid Monitoring



MDS is used in the monitoring system

May 12-15, 2003

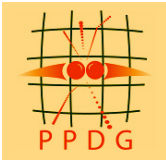
Lee Lueking, EDG Int. Proj. Conf.

23

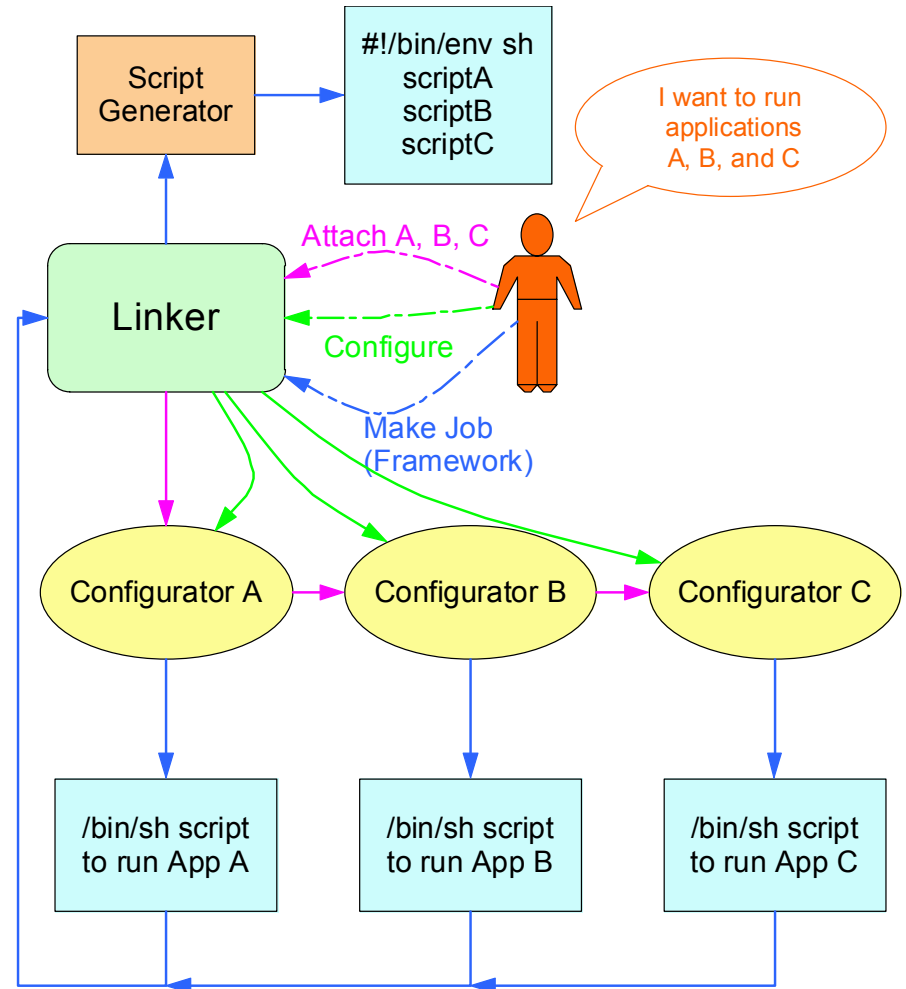




# Meta Systems



- MCRunJob approach by CMS and DØ production teams
- Framework for dealing with multiple grid resources and testbeds (EDG, IGT)





# DO JIM Deployment



- **A site can join SAM-Grid with combinations of services:**
  - **Monitoring, and/or**
  - **Execution, and/or**
  - **Submission**
- **May 2003: Expect 5 initial execution sites for SAMGrid deployment, and 20 submission sites.**
- **Summer 2003: Continue to add execution and submission sites.**
- **Grow to dozens execution and hundreds of submission sites over next year(s).**
- **Use grid middleware for job submission within a site too!**
  - **Administrators will have general ways of managing resources.**
  - **Users will use common tools for submitting and monitoring jobs everywhere.**





# What's Next for SAM-Grid?

After JIM version 1

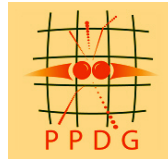


- **Improve scheduling jobs and decision making.**
- **Improved monitoring, more comprehensive, easier to navigate.**
- **Execution of structured jobs**
- **Simplifying packaging and deployment. Extend the configuration and advertising features of the uniform framework built for JIM that employs XML.**
- **CDF is adopting SAM and SAM-Grid for their Data Handling and Job Submission.**
- **Co-existence and Interoperability with other Grids**
  - **Moving to Web services, Globus V3, and all the good things OGSA will provide. In particular, interoperability by expressing SAM and JIM as a collection of services, and mixing and matching with other Grids**
  - **Work with EDG and LCG to move in common directions**





# Run II plans to use the Virtual Data Toolkit

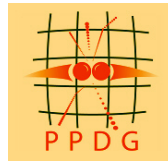


- JIM is using advanced version of Condor-G/Condor - actually driving the requirements. Capabilities available in VDT 1.1.8 and beyond.
- D0 uses very few VDT packages- Globus GSI, GridFTP, MDS and Condor.
- JIM ups/upd packaging includes configuration information to save local site managers effort. Distribution and configuration tailored for existing/long legacy D0 systems.
- Plans to work with VDT such that D0-JIM will use VDT in the next six months.
- ==>> VDT versions are currently being tailored for each application community. This cannot continue. We - D0, US CMS, PPDG, FNAL, etc.- will work with the VDT team and the LCG to define how VDT versions should be
  - Constructed and Versioned
  - Configured
  - Distributed to the various application communities
  - Requirements and scheduled for releases.

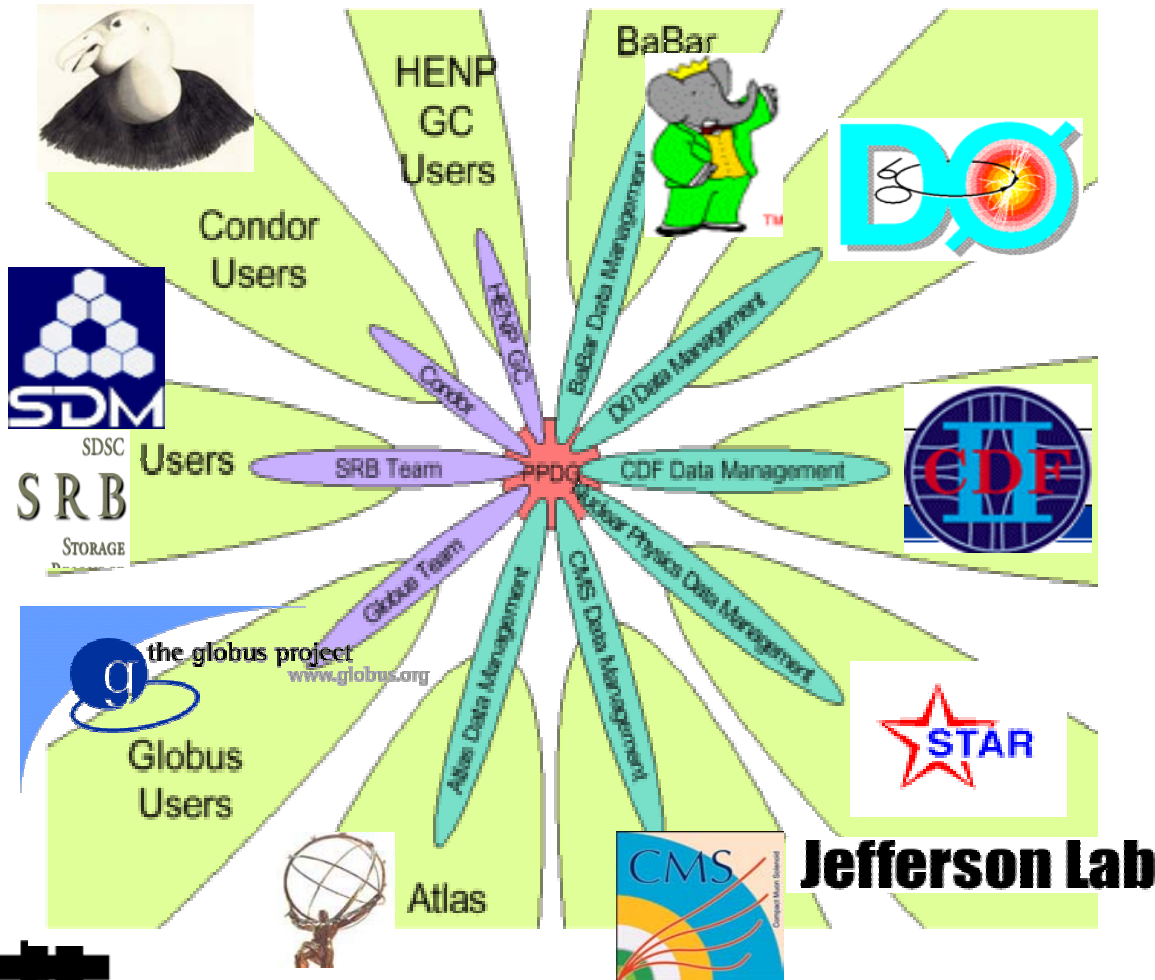




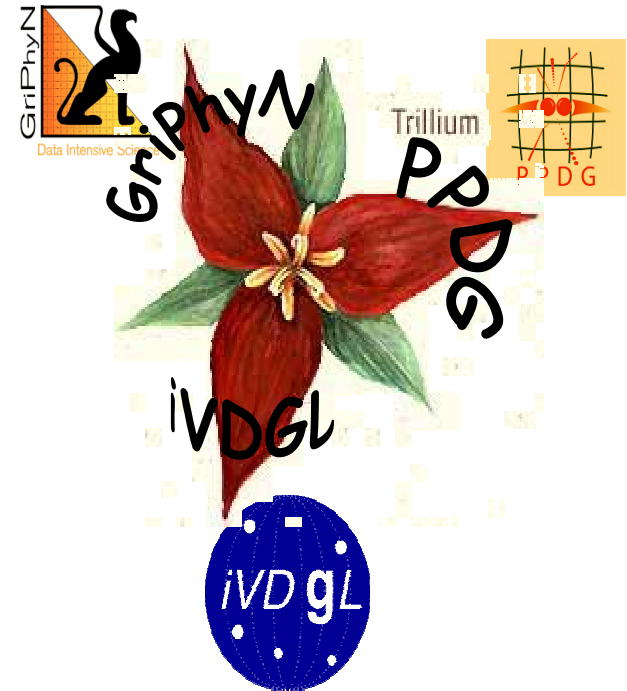
# Projects Rich in Collaboration



## PPDG



## Trillium

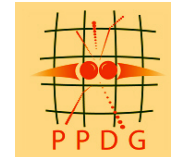


May 12-15, 2003

Lee Lueking, EDG Int. Proj. Conf.

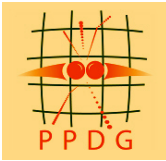


## Collaboration between Run 2 and US CMS Computing at Fermilab



- D0, CDF, and CMS are all using Dcache and Enstore storage management systems.
- Grid VO management - joint US-CMS, iVDGL, INFN-VOMS, (LCG?) project is underway
  - <http://www.uscms.org/s&c/VO/meeting/meet.html>
  - There is a commitment from the RUN II Experiments to collaborate on with this effort in near future.
- (mc)Runjob scripts - joint work on core framework between CMS and Run II experiments has been proposed.
- Distributed and Grid accessible databases and applications are a common need.
- As part of PPDG we expect to collaborate on future projects such as Troubleshooting Pilots (end to end error handling and diagnosis).
- Common infrastructure in Computing Division for system and core service support etc. ties us together.





# Regional Computing Approach



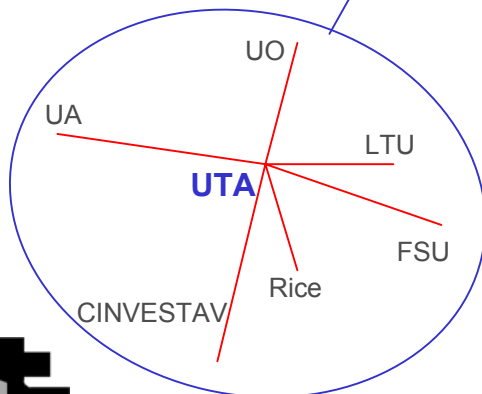
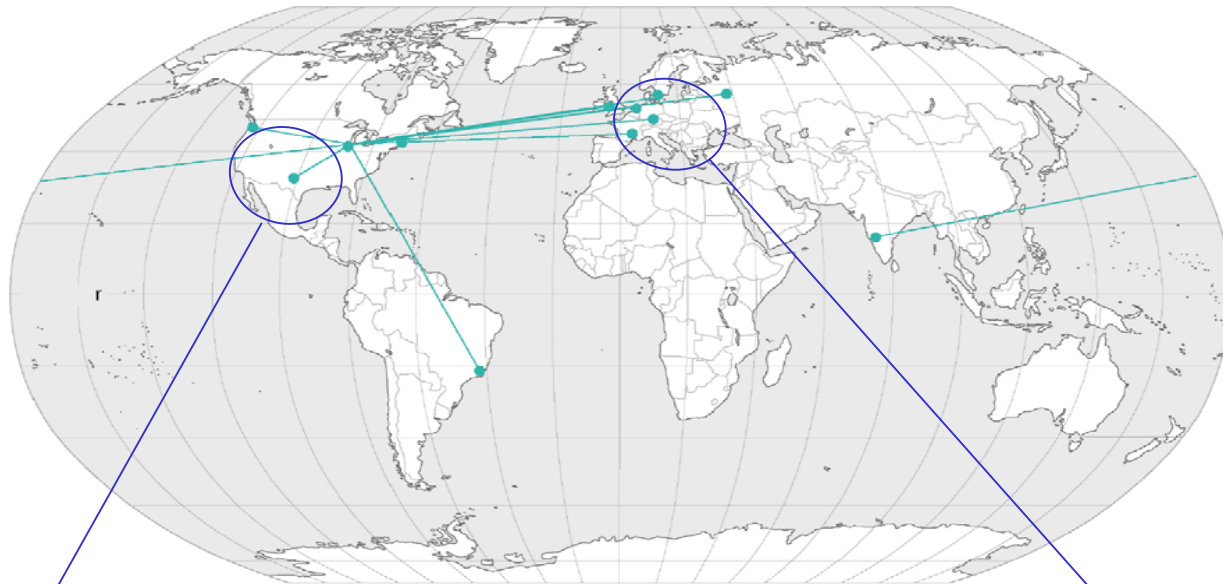
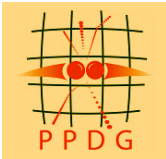
May 12-15, 2003

Lee Lueking, EDG Int. Proj. Conf.

30



# DØ Regional Model



Centers also in the UK and France

UK: Lancaster, Manchester, Imperial College, RAL

France: CCin2p3, CEA-Saclay, CPPM Marseille, IPNL-Lyon, IRES-Strasbourg, ISN-Grenoble, LAL-Orsay, LPNHE-Paris



May 12-15, 2003

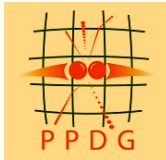
Lee Lueking, EDG Int. Proj. Conf.

31





# Regional Analysis Centers (RAC) Functionality

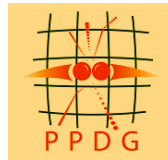


- **Preemptive caching**
  - **Coordinated globally**
    - All DSTs on disk at the sum of all RAC's
    - All TMB files on disk at all RACs, to support mining needs of the region
  - **Coordinated regionally**
    - Other formats on disk:  
Derived formats & Monte Carlo data
- **On-demand SAM cache: ~10% of total disk cache**
- **Archival storage (tape - for now)**
  - Selected MC samples
  - Secondary Data as needed
- **CPU capability**
  - supporting analysis, first in its own region
  - For re-reconstruction
  - MC production
  - General purpose DØ analysis needs
- **Network to support intra-regional, FNAL-region, and inter-RAC connectivity**

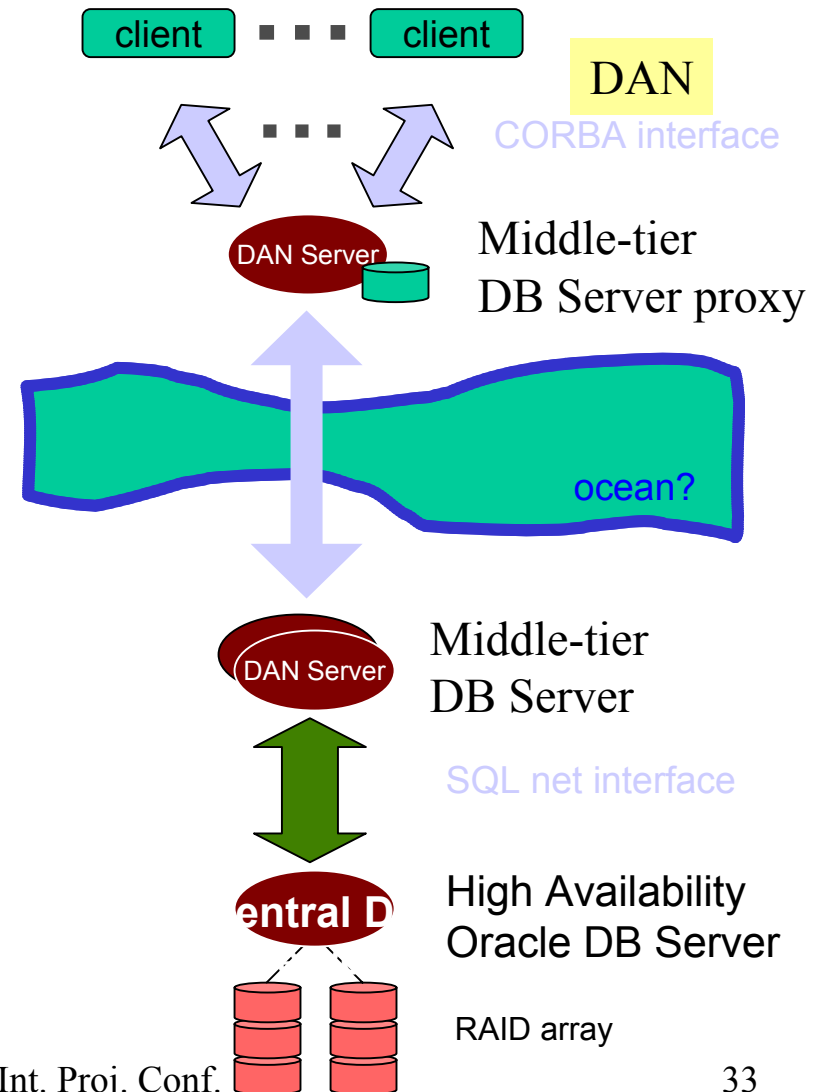
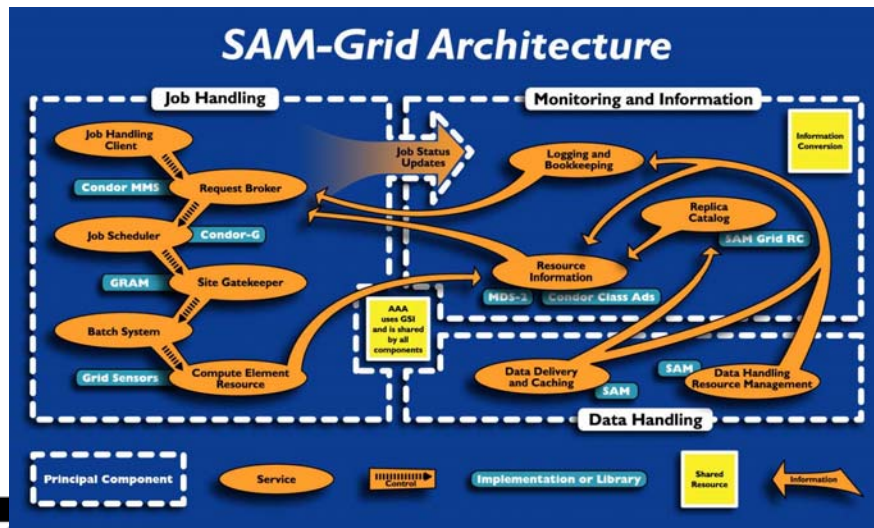




# Required RAC Server Infrastructure

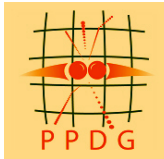


- SAM-Grid Gateway machine
- Oracle database access servers
  - Provided via middle tier server (DAN)
  - DAN = Database Access Network
- Accommodate realities like:
  - Policies and culture for each center
  - Sharing with other organizations
  - Firewalls, private networks, et cetera





# Summary of Current & Soon-to-be RACs



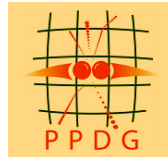
Regional Centers	Institutions within Region	CPU ΣHz (Total*)	Disk (Total*)	Archive (Total*)	Schedule
GridKa @FZK	Aachen, Bonn, Freiburg, Mainz, Munich, Wuppertal,	52 GHz (518 GHz)	<div>Total Remote CPU 360 GHz (1850 GHz)</div>		Established as RAC
SAR @U (Southern)	<div>Total need for Beginning of 2004 ~4500 GHz</div>	160 GHz (320 GHz)			Summer 2003
UK @tbd		46 GHz (556 GHz)			Active, MC production
IN2P3 @Lyon		100 GHz			Active, MC production
	LPNHE-Paris		<div>FNAL CPU 1800 GHz</div>		
DØ @FNAL (Northern US)	Farm, cab, clued0, Central-analysis	1800 GHz			Established as CAC

\*Numbers in () represent totals for the center or region, other numbers are DØ's current allocation.





# Data Model

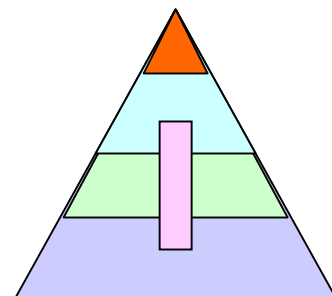


## Fraction of Data Stored

Data Tier	Size/event (kB)	FNAL Tape	FNAL Disk	Remote Tape	Remote Disk
RAW	250	1	0.1	0	0
Reconstructed	50	0.1	0.01	0.001	0.005
DST	15	1	0.1	0.1	0.1
Thumbnail	10	4	1	1	2
Derived Data	10	4	1	1	1
MC D0Gstar	700	0	0	0	0
MC D0Sim	300	0	0	0	0
MC DST	40	1	0.025	0.025	0.05
MC TMB	20	1	1	0	0.1
MC PMCS	20	1	1	0	0.1
MC root-tuple	20	1	0	0.1	0
Totals RIIa ('01-'04)/ RIIb ('05-'08)		1.5PB/ 8 PB	60TB/ 800 TB	~50TB	~50TB

per Region

## Data Tier Hierarchy



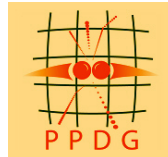
▲ Metadata  
~0.5TB/year

Numbers are  
rough estimates

the cpb model presumes:  
25Hz rate to tape, Run IIa  
50Hz rate to tape, Run IIb  
events 25% larger, Run IIb



# Challenges

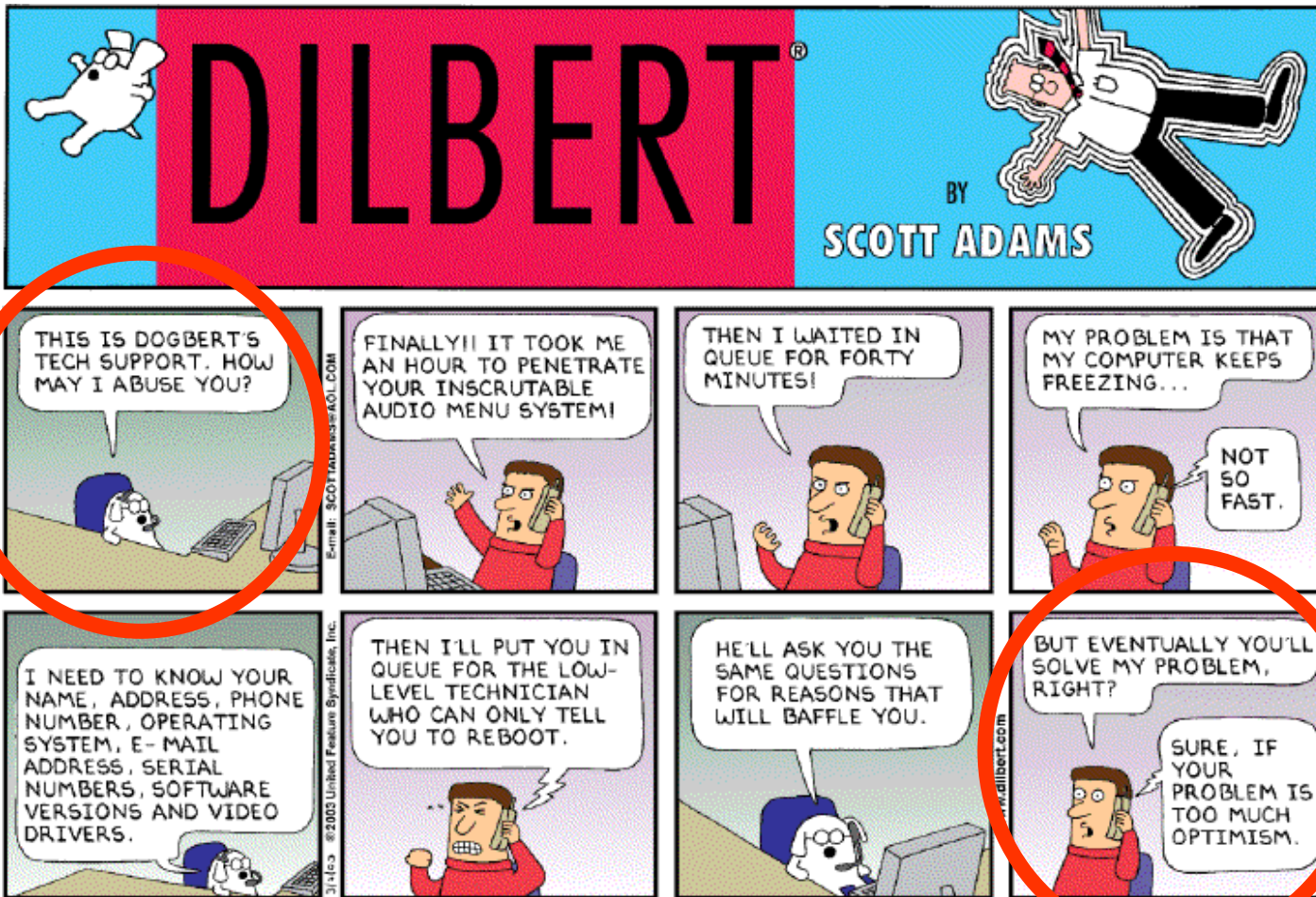
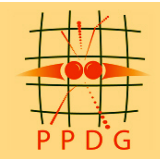


- Operation and Support
  - Ongoing shift support: 24/7 “helpdesk” shifters (trained physicists)
  - SAM-Grid station administrators: Expertise based on experience installing and maintaining the system
  - Grid Technical Team: Experts in SAM-Grid, DØ software + technical experts from each RAC.
  - Hardware and system support provided by centers
- Production certification
  - All DØ MC, reconstruction, and analysis code releases have to be certified
- Special requirements for certain RAC’s
  - Forces customization of infrastructure
  - Introduces deployment delays
- Security issues, grid certificates, firewalls, site policies.





# Operations



Copyright © 2003 United Feature Syndicate, Inc.

## Expectation Management

May 12-15, 2003

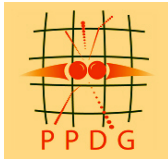
Lee Lueking, EDG Int. Proj. Conf.

37



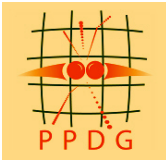


# Summary



- The DØ Experiment is moving toward exciting Physics results in the coming years.
- The Data Management software is stable and provides reliable data delivery and management to production systems worldwide.
- SAM-Grid is using standard Grid middleware to enable complete Grid functionality. This is rich in collaboration with Computer Scientists and other Grid efforts.
- DØ will rely heavily on remote computing resources to accomplish its Physics goals





Thank You



May 12-15, 2003

Lee Lueking, EDG Int. Proj. Conf.

39